

WiSPA: A new approach for dealing with widespread parasitism

# WiSPA: A new approach for dealing with widespread parasitism

BENJAMIN DRINKWATER<sup>1</sup>, ANGELA QIAO<sup>1</sup>, AND MICHAEL A. CHARLESTON<sup>1 2</sup>

<sup>1</sup>*School of Information Technologies, University of Sydney, NSW, 2006, Australia;*

<sup>2</sup>*School of Physical Sciences, University Of Tasmania, TAS, 7005, Australia;*

**Corresponding author:** Benjamin Drinkwater, School of Information Technologies (J12), University of Sydney, NSW, 2006, Australia; E-mail: benjamin.drinkwater@sydney.edu.au

*Abstract.*— Traditionally, studies of coevolving systems have considered cases where a parasite may inhabit only a single host. The case where a parasite may infect many hosts, *widespread parasitism*, has until recently gained little traction. This is due in part to the computational complexity involved in reconstructing the coevolutionary histories where parasites may infect only a single host, which is NP-Hard. Allowing parasites to inhabit more than one host has been seen to only further compound this computationally intractable problem. Recently however, well-established algorithms for estimating the problem instance where a parasite may infect only a single host have been extended to handle widespread parasites. Although this has offered significant progress, it has been noted that these algorithms poorly handle parasites that inhabit phylogenetically distant hosts.

In this work we extend these previous algorithms to handle cases where parasites inhabit phylogenetically distant hosts using an additional evolutionary event which we call *spread*. Our new framework is shown to infer significantly more congruent coevolutionary histories compared to existing methods over both synthetic and biological data sets. We then apply the newly proposed algorithm, which we call WiSPA (WideSpread Parasitism Analyser), to the well studied coevolutionary system of *Primates* and *Enterobius* (pinworms), where existing methods have been unable to reconcile the widespread parasitism present without permitting additional divergence events. Using WiSPA and the new biological event, spread, we provide the first statistically significant coevolutionary hypothesis for this system.

(Keywords: Coevolution, Phylogeny, Widespread Parasitism, NP-Hard )

Coevolutionary research has long focused on the area of parasitism due to the health risks which parasites pose to the human population (Charleston and Perkins 2006). Parasites and the associations they form with their hosts have been responsible for a number of the worst emerging diseases impacting global health today, including *Ebola* (Peterson et al. 2004), HIV (Siddall 1997), and malaria (Mu et al. 2005). Further research into the field of coevolution aims to uncover the deep coevolutionary associations formed by parasitic behaviour, to provide further insights into these deadly diseases (Charleston and Galvani 2006).

We often define coevolutionary systems in terms of an independent phylogeny and a corresponding dependent phylogeny which have formed a macro-scale coevolutionary bond.

One approach that is often applied to evaluating such evolutionary relationships is the field of *cophylogenetics*, which provides a framework to evaluate whether evolutionary histories have coevolved or have evolved independently (Charleston 2003).

As a result of host–parasite systems’ long association with the field of cophylogenetics, coevolutionary systems often describe the independent and dependent phylogenies as the *host* ( $H$ ) and *parasite* ( $P$ ) respectively. Cophylogenetic analysis, however, can be applied to all forms of coevolutionary dependence including: biogeography (Toit et al. 2013), host–pathogen systems (Mu et al. 2005), genes and the species that house them (Page and Charleston 1997), plant–insect interactions (Gómez-Acevedo et al. 2010), plant–fungi dynamics (Refrégier et al. 2008), host–parasitoid relationships (Stireman 2005), and Batesian and Müllerian mimicry between species (Ceccarelli and Crozier 2007; Cuthill and Charleston 2012).

The coevolutionary interactions between  $P$  and  $H$  are represented by the associations ( $\varphi$ ) between their leaves, based on evidence of parasites inhabiting or infecting their host(s). These associations can be used to infer the level of host specificity of the parasite species with respect to its host(s) (Poulin 2011). Within this context, high host specificity is the case where a particular parasite infects a single host species, while low host specificity is the case where a parasite may infect many host species.

Coevolutionary analysis of systems with high host specificity focuses on the reconstruction of the parasite’s evolutionary history with respect to the host, which is known as a *cophylogeny mapping*. When recovering a map ( $\Phi$ ) using cophylogeny mapping, the aim is to recover the most congruent solution with a minimum total event cost, while ensuring the associations are conserved using the four known coevolutionary events, *codivergence*, *duplication*, *host switch* and *loss* (Ronquist 1995).

A codivergence event is a concurrent divergence of both the host and parasite lineages. A high concentration of codivergence events leads to an increase in the level of

congruence between  $P$  and  $H$ , and is therefore a strong indicator of coevolution (Page 2002). A duplication event is an independent divergence of the parasite where both new lineages continue to track the host (Tuller et al. 2010). A host switch event is an independent divergence of the parasite where one parasite shifts from the initial host lineage (take-off edge) to a new lineage (landing edge) in the host, while the second parasite continues to track the host (Kim et al. 1985). We call these three events *divergence events*, as they consider all cases of divergence in the parasite’s coevolutionary history. By contrast, loss arises from three indistinguishable processes: lineage sorting (or “missing the boat”), extinction, or sampling failure. As these processes all produce the same effect we represent these as a *loss* event (Paterson et al. 2003). We refer to the problem of reconstructing a map using only these four events as the *restricted cophylogeny reconstruction problem*.

Methods for recovering maps have mainly focused on the restricted cophylogeny reconstruction problem. This is due in part to the initial set of biological events being unable to reconstruct the evolutionary history of parasites with low host specificity, along with the hypothesis that coevolution only occurs in systems with a one-to-one association between parasites and their hosts (Poulin 2011). This hypothesis, however, only considers a select set of coevolving systems and precludes many observed coevolutionary systems where the parasites maintain low host specificity as an evolutionary advantage. In a comprehensive study of plant–insect interactions, Nosil and Mooers (2005) demonstrated that while insects often form exclusive associations with their hosts, this is not always the case. Butterflies and bark beetles were shown to often be associated with many host plant species. This case is not unusual, with Stireman’s (2005) study of endoparasitoids and their *Tachinidae* (fly) hosts also demonstrating the evolutionary advantage of low host specificity. These results affirm that ongoing cophylogeny mapping modelling must consider the general case where parasites are permitted to inhabit more than one host

(*widespread parasitism*), to accurately model all coevolutionary interrelationships.

As described above, modelling widespread parasitism using cophylogeny mapping requires additional biological events beyond the original four events derived by Ronquist (1995). Currently, failure-to-diverge is the only event which has successfully been applied to handle widespread parasitic events within a cophylogeny mapping framework. It is defined as the case where parasites maintain their ability to inhabit both hosts following a host divergence event, without the need for a divergence of the parasite lineage (Johnson et al. 2003). Failure-to-diverge is the case where there is an interruption of the gene-flow between the host species while there remains gene flow within the parasite population (Poulin 2011). A case where failure-to-diverge and the full set of divergence events are required to reconstruct a cophylogenetic history can be seen in Figure 1. We will here refer to events which are used to describe widespread parasite coevolution, such as failure-to-diverge, as *widespread events*. This is to differentiate such events from divergence and loss events.

The failure-to-diverge event allows for the recovery of solutions for all conceivable cases of widespread parasitism for cophylogenetic reconstructions. However, these solutions may have a high number of loss events when widespread parasites inhabit phylogenetically distant leaves in  $H$ . This is due to the limitation that a failure-to-diverge event occurs at the most recent common ancestor of the pair of inhabited host leaves (Banks and Paterson 2005), after which many loss events must be inferred to account for the observed parasite distribution.

Cophylogenetic reconstructions are evaluated using an event cost, similar to that of a parsimony score in phylogenetic reconstructions (Charleston 2002). Reconstructing the minimum cost map requires that each divergence event, widespread event, and loss event be assigned a penalty cost. The set of costs for each event may be defined as a vector  $V = (C, D, W, L, F)$  where  $C, D, W, L, F$  represent the associative costs for each codivergence, duplication, host switch, loss, and failure-to-diverge respectively. The

resultant map cost  $E$  can then be derived as:

$$E = \alpha C + \beta D + \gamma W + \delta L + \epsilon F \quad (1)$$

where  $\alpha$ ,  $\beta$ ,  $\gamma$ ,  $\delta$ ,  $\epsilon$  represent the number of events for codivergence, duplication, host switch, loss, and failure-to-diverge respectively (Drinkwater and Charleston 2014a).

Cophylogeny mapping algorithms aim to map  $P$  into  $H$ , where the number of codivergence events is maximised and the map cost  $E$  is minimised (Charleston and Libeskind-Hadas 2014). Although such cost schemes may not evaluate coevolutionary scenarios exactly, particularly modelling preferential host switching (Charleston and Robertson 2002), this technique has been used to evaluate a large number of coevolutionary systems (Page 2002; Page et al. 2004; Jackson and Charleston 2004; Cruaud et al. 2012; Rivera-Parra et al. 2015).

Parsimony and event-based methodologies are often seen as less preferable to maximum likelihood methodologies. This is due to parsimony methods often relying on arbitrarily chosen cost schemes. While such a reliance is a limiting factor, parsimony and event based methods may be used to reconstruct the most likely evolutionary history by assigning the negative log likelihood probabilities for each evolutionary event as the associated penalty cost for each event (Drinkwater and Charleston 2016). As a result there is a strong driver for fast mapping methods which can then be integrated into a coevolutionary likelihood framework (Charleston 2003), to complement maximum likelihood techniques (Baudet et al. 2015).

Unfortunately recovering the minimum cost map is known to be NP-Hard (Ovadia et al. 2011). This computational intractability is due to the exponential number of host switch locations that can arise due to the variable order of the internal nodes in the host tree (Doyon et al. 2010), and the exponential number of internal node orderings (Conow

et al. 2010).

To mitigate this computational intractability two unique heuristics have been proposed. The first approach ignores the relative ordering of the internal nodes in the parasite phylogeny, This may lead to the order of evolutionary events, as defined by a reconciled map, contradicting the order of evolutionary events as defined by the parasite phylogeny (Doyon et al. 2010). Such a map is often referred to as *biologically infeasible* or time-inconsistent (Doyon et al. 2011). This approach has been applied in various methods (Merkle and Middendorf 2005; Merkle et al. 2010; Yodpinyanee et al. 2011), with the fastest known algorithm to date running in  $O(n^2)$  (Bansal et al. 2012).

To guarantee that solutions are time-consistent two properties must be ensured. A host switch's take-off and landing edges must lie in the same time interval, which is an overlapping interval based on the edges' distances from the root of  $H$ , and must maintain the partial ordering of  $P$  (Conow et al. 2010), as mentioned above. This requires that the relative order of the parasite phylogeny must be fixed, which has led to the second heuristic which fixes the internal node ordering of the host phylogeny. If the internal node ordering is fixed it is possible to solve the cophylogeny reconstruction problem in polynomial time. This simplified problem, often referred to as the dated tree reconciliation problem (Drinkwater and Charleston 2015a), has been applied within a number of algorithms (Doyon et al. 2011; Conow et al. 2010; Drinkwater and Charleston 2015b), with the fastest proposed to date running in  $O(n^2 \log n)$  (Bansal et al. 2012).

While the aim of the cophylogeny reconstruction problem is to recover the minimum cost map in terms of,  $E$ , it is often valuable to infer the significance of the resultant map. In particular it is valuable to identify if a resultant map provides a statistically significant signal that the apparent congruence is unlikely to have occurred simply by chance. Previously, analysis has applied a series of Bernoulli trials to analyse the significance of an inferred map, such as the analysis of pocket-gophers and their parasitic chewing lice by

Page in 1994a. This is also the premise of the statistical evaluation tool Parafit (Legendre et al. 2002) which randomises the associations or the parasite tree to identify whether the degree of congruence noted between the host and parasite tree could have occurred simply by chance. This process can be replicated when using cophylogeny mapping, by producing randomised permutations of the initial tanglegram (either randomising the associations or the parasite tree), and computing the cost of the optimal map for this randomised instance (Page 1994a,b). If the initial tanglegram has a mapping cost,  $E$ , which is less than the randomised permutations in at least 95% of cases, then we may reject the null hypothesis that there is no significance between the independent and dependent phylogenetic trees. This feature is currently integrated into the most recent implementation of the Jane software tool (Conow et al. 2010).

Three recent software tools which have been designed to recover maps where widespread parasites are considered and using the failure-to-diverge widespread evolutionary event, are CoRe-PA (Merkle et al. 2010), Jane (Conow et al. 2010) and CoRe-ILP (Wieseke et al. 2015). CoRe-PA solves the cophylogeny reconstruction problem in polynomial time by relaxing the internal node ordering of the host tree. This approach, although potentially recovering solutions in quadratic time (Yodpinyanee et al. 2011), may recover solutions which are time-inconsistent (Doyon et al. 2011). Jane, in contrast, fixes the internal node order in the host tree and solves this instance using dynamic programming (Libeskind-Hadas and Charleston 2009). This approach guarantees that solutions are biologically feasible. As there are an exponential number of possible fixed node orderings, most techniques applying this approach leverage a genetic algorithm to recover the best possible solutions in a fixed period of time. Finally, CoRe-ILP applies an integer linear programming algorithm to solve the problem of maximising the total number of codivergence events within the reconciled map, which its developers have shown provides a robust estimation of the harder problem of finding the minimum cost map (Wieseke et al.



2015).

While each approach handles the computational intractability of recovering the divergence events in significantly different ways, all apply a common approach for recovering widespread events, where each failure-to-diverge event occurs at the most recent common ancestor of the parasite’s host, guaranteeing solutions can be recovered in all cases. This approach while offering researchers the first set of tools for inferring coevolutionary systems which include widespread parasites, often infers maps with a high number of loss events polluting the coevolutionary signal.

Our work aims to expand on this research by constructing a new methodology which solves the cophylogeny reconstruction problem with widespread parasites, the *widespread parasite problem*, which is able to overcome the high costs that are often associated with failure-to-diverge. As a result, this method will decrease the overall parsimony score, while potentially increasing the number of codivergence events within the reconciled map.

## METHODOLOGY

### *Reintroducing an additional biological event for coevolutionary analysis*

This work reintroduces an additional evolutionary event for inferring relationships of coevolutionary systems where widespread parasites are permitted, which we call *spread*. We propose that existing frameworks be updated to include spread as an additional widespread parasite event, to work in conjunction with failure-to-diverge. The inclusion of spread aims to more accurately reconcile the widespread parasites’ coevolutionary histories with respect to their hosts, by mitigating the high number of loss events which are associated with reconstructions that exclusively use failure-to-diverge, such as cases where parasites *do not* inhabit closely related hosts.

The spread event was first applied by Brooks, (1991), to reconcile widespread parasites within the Brooks Parsimony Analysis framework, and was later proposed by Siddall and Perkins, (2003), as an additional widespread evolutionary event to be integrated within TreeMap (Charleston 2012). In both cases, however, neither of these proposed models have been implemented in part due to the additional computational complexity that their inclusion can give rise to.

The spread event is derived from a number of observed parasitic systems, such as the behaviour of chewing lice which infect their penguins hosts (?). It has been observed that a number of lice species switch between their penguin hosts at shared breeding grounds, however, this cannot be modelled using a host switch, as there is no divergence in the parasite lineage, nor can this be considered a failure-to-diverge event as there is no evidence to support that the gene flow has been maintained for the chewing lice species. Rather, the lice species have recently spread to new hosts based on new opportunities that are presented.

This “spreading” behaviour has also been observed in lab experiments between nematodes and their *Drosophila* fly hosts (Jaenike and Dombeck 1998). Nematodes’ general purpose genotypes allow each individual species to infect a high number of host species, allowing nematodes to infect distantly related *Drosophila* hosts which they would not be expected to encounter in nature. The infection of *Drosophila* does not require any evolutionary changes so this cannot be modelled correctly using a host switch event, nor can it be described using failure-to-diverge as the nematodes have not coexisted with their new hosts, and therefore this phenomenon requires an additional biological event to model this observed behaviour, which spread successfully achieves.

The spread event also complements the theory of low host specificity which asserts species evolve specific mechanisms which allow them to inhabit multiple hosts where the parasites are less vulnerable to the evolutionary changes of a specific host species. Further,

spread often provides more parsimonious solutions for widespread parasitism. Consider the coevolutionary system in Figure 2 (left). In the first reconstruction, Figure 2 (center), the parasite has had  $O(n)$  opportunities to infect a new host species but failed to do so in all cases. This is highly unlikely compared to the alternate reconstruction using spread, where no loss events occur and the parasite simply infects a new host species based on new opportunities, such as the introduction of an infected host species ( $A$ ) into host species ( $B$ ) natural environment.

The alternate map which uses spread, Figure 2 (right), is significantly more parsimonious for cases where the spread event is assigned a penalty cost similar to that of failure-to-diverge. In fact, spread would need to be assigned a cost  $n$  times that of a loss event for the solution to be considered more expensive. Therefore, as this biological event describes observed behaviour in nature and also allows for potentially more parsimonious maps, we argue that spread should be integrated into existing algorithms which aim to infer systems which present widespread parasites. This is in line with previous assertions made by Brooks, (1991), Page, (1994a), and Siddall and Perkins, (2003).

While often producing significantly cheaper solutions, spread may not always be possible as it is reliant on host species collocation to permit the occurrence of a spread event similar to host switch events (Clayton et al. 2004). Further research needs to be undertaken on how to model this and complements the existing field of research into preferential host switching (Charleston and Robertson 2002; Cuthill and Charleston 2013). We, however, do not consider this constraint herein, and assume spread is permissible between all hosts, as a means to present this evolutionary event's value to widespread parasitism analysis.

Formally we define the spread event as a parasite lineage that due to new opportunities infects a new host lineage while maintaining its infection of its current host lineage. This event therefore consists of a shift of a subset of the parasite lineage from the

initial host (the take-off edge) to a new host (the landing edge), occurring at some point after the host lineages have diverged.

This definition is derived from the existing definition of the host switch event with which spread shares a number of common traits. Both events require the internal node ordering of the *host* phylogeny to be fixed, to ensure that the resultant map is time consistent, and both events require that the take-off and landing edges share a common timing interval (Conow et al. 2010). Spread events, unlike host switch events however, do not consist of a bifurcation, and as a result are not dependent on the internal node ordering of the *parasite* phylogeny, which results in spread being a more generalised version of a host switch event.

With the addition of the spread event we are required to update the cost vector  $V$  to include  $S$ , the cost of a spread event, along with updating the objective function  $E$  as follows:

$$E = \alpha C + \beta D + \gamma W + \delta L + \epsilon F + \zeta S \quad (2)$$

where  $\zeta$  represents the number of spread events in the resultant map,  $\Phi$ . It is important to note that even with the addition of spread as an additional evolutionary event, the total number of widespread events in Equation (2) ( $\epsilon + \zeta$ ) is equal to the number of failure-to-diverge events in Equation (1) ( $\epsilon$ ).

Using this new formulation of the objective function  $E$ , we derive a polynomially bounded algorithm to solve the cophylogeny reconstruction problem where widespread parasites are permitted (the widespread parasites problem), where the internal nodes in the host phylogeny are fixed. The proposed method extends the Improved Node Mapping algorithm (Drinkwater and Charleston 2014a, 2015a), to recover solutions to the widespread parasite problem using both spread and failure-to-diverge. The described

methodology, however, is designed so that it can be integrated into other mapping algorithms which leverage a fixed internal node ordering, such as Edge Mapping (Yodpinyanee et al. 2011) and Slicing (Doyon et al. 2010). This method is then integrated into an existing metaheuristic framework similar to that implemented in Jane (Conow et al. 2010), which allows for this method to provide robust estimations for the widespread parasites problem in a reasonable period of time.

### *The order of evolutionary events*

Along with integrating both spread and failure-to-diverge within a common framework, our model aims to provide additional flexibility when inferring the position of a widespread event within the reconciled map. Current state of the art algorithms such as Edge Mapping applied in Jane, provide strict bounds on the position where a failure-to-diverge event may occur. These bounds only allow for a subset of the total number of mapping locations to be considered prior to the widespread event. For example consider the tanglegram in Figure 3 (left) which includes a single widespread parasite. The minimum cost map inferred by Jane, Figure 3 (right), for this specific instance includes 2 failure-to-diverge events, 1 host switch event and 1 loss event.

There is an alternate reconstruction for this system, however, where the minimum cost map contains 2 failure-to-diverge events and 1 codivergence event, Figure 3 (centre). Under all previously published cost schemes this map is considered more parsimonious. Jane is unable to reconstruct this specific map, however, as its algorithm enforces constraints on the number of locations where divergence events may be placed following a set of widespread events. This bound is appropriate as it does allow for a faster running time, however, this bound in cases such as this may give rise to reconciliations which are less parsimonious.

As computational power continues to become faster and cheaper, it is important to consider alternate algorithms which, while potentially less efficient, may provide more parsimonious solutions to the widespread parasite problem. This is the concept which is explored herein, where our proposed framework permits divergence events to occur at all feasible positions prior to and following a set of widespread events. This will increase the asymptotic complexity relative to Jane, in the hope of providing a more parsimonious reconciliation for the resultant maps.

We will show that by increasing the asymptotic complexity by a factor of  $n$  that it is possible to provide a solution to the widespread parasite problem which considers both failure-to-diverge and spread, and provides a significant accuracy improvement which is representative of one of the largest single improvements offered by a coevolutionary analysis technique since Charleston (1998) proposed the Jungle data structure.

### *Integrating Widespread Events into Improved Node Mapping*

In this section we introduce a series of amendments which when applied to the Improved Node Mapping algorithm allows for both failure-to-diverge and spread events to be recovered optimally when reconciling a pair of phylogenetic trees. Prior implementations of node mapping by Libeskind-Hadas and Charleston (2009), and Drinkwater and Charleston (2014a; 2015b) have only considered the case where a parasite may inhabit a single host. The amendment described herein not only updates the Improved Node Mapping algorithm to support widespread events, but also resolves the problems associated with algorithms such as Jane which were discussed in the previous section.

The updated version of the Improved Node Mapping algorithm which we will refer to as WiSPA (WideSpread Parasitism Analyser) can be more easily described as a two step process. The first reconciles all optimal widespread events based on an event costs vector,  $V$ . This is a reconciliation step which recovers all feasible widespread events, where the

second step recovers the optimal set of divergence events using the previously derived set of widespread events. By handling these two complex sets of operations in series it is possible to ensure that a polynomially bound algorithm may be derived for solving the widespread parasite problem, where the internal node ordering of the host phylogeny is fixed.

Our proposed algorithm reconciles the set of optimal widespread events by constructing a set of widespread association trees, a process which is derived from an earlier method proposed by Page (1994b). These association trees are then leveraged to recover the optimal set of widespread events, mirroring much of the work proposed by both Page (1994b) and Siddall and Perkins (2003). Unlike their previous attempts to solve the widespread parasitism problem which applied a greedy algorithm, our approach applies a dynamic programming algorithm to ensure that all feasible states may be considered, avoiding the potential problems that may arise due to local minima or excluding large subsets of the problem space.

### *Reconstructing Widespread Associations as Trees*

To reconcile the set of widespread events for each widespread parasite ( $p_i$ ), we propose a method which translates the set of widespread associations for the parasite node  $p_i$  to a bifurcating tree. A similar model was first used by Page in 1994b to reconcile the widespread parasitism identified in the pocket gopher chewing lice coevolutionary system introduced by Hafner and Nadler, (1988).

The constructed trees which are referred to herein as Association Trees ( $a_i$ ), are a set of trees  $A = (a_1 \dots a_n)$ , which may be used to infer the optimal set of widespread events where we prove that:

**Lemma 1.** *An association tree ( $a_i$ ) is a bifurcating tree constructed based on the associations,  $\varphi$ , present for the parasite leaf node  $p_i$  which mirrors the topology of  $H$ , such that  $a_i$  may infer the maximum number of widespread events.*

*Proof.* Consider the parasite leaf node  $p_i$  with  $k$  widespread associations. The maximum number of possible widespread events is the case where  $k$  failure-to-diverge events may be recovered. This is because for all cases it is possible to recover  $k$  spread events for all trees, due to the construction of the host tree, such that all leaves share a common timing interval (the present) (Conow et al. 2010). Therefore an association tree,  $a_i$ , which maximises the number of failure-to-diverge events will maximise the total number of possible widespread events.

A mirrored tree constructed in line with Fahrenholz’s (1913) Rule will always permit  $k$  failure-to-diverge events, as each internal node in the mirrored tree corresponds to an internal node in the host tree (Fahrenholz 1913; Paterson and Banks 2001). Therefore if we construct  $a_i$  for  $p_i$  which mirrors  $H$  based on the associations  $\varphi$ , then we will maximise the number of possible widespread events which are able to be recovered using the association tree. □

By maximising the number of possible events recovered, we ensure that the optimal set of widespread events may be inferred. This is due to the order of widespread events being unbounded, as widespread events are not dependent on the internal order of  $P$ . Therefore, this approach while ensuring that an optimal set of widespread events is recovered, does not guarantee that the order of events inferred is correct, as there is no information in the initial problem instance to provide such an inference. Further, information about the problem instance would be required to infer the order of widespread events, such as the geographical history of both host and parasite. This, along with the consideration of preferential spread events, is a topic to be considered in later revisions of the WISPA algorithm.

In order to construct the association trees in line with Fahrenholz’s (1913) Rule, we find the unique subtree where each leaf in the association tree is associated with one of the



initial widespread associations. The recovery of an association tree can therefore be reduced to the problem of recovering the homeomorphic subgraph of  $H$  for the leaves inhabited by the widespread parasite  $p_i$  (Lozano et al. 2007).

To construct the homeomorphic subgraph we apply the pruning algorithm described in detail by Lozano et al. (2007) which creates a copy of  $H$  where only the host leaves inhabited by  $p_i$  are retained. This algorithm is applied for each widespread parasite which gives rise to the set  $A = (a_1 \dots a_n)$ .

The associations trees  $A = (a_1 \dots a_n)$  mirror  $H$  based on each parasite's widespread associations and therefore each leaf in the association tree  $a_i$  has a one-to-one association with a leaf in  $H$ , such that each leaf node in the association tree  $a_i$  maps to a unique leaf node in  $H$ . This property is not one that is imposed on a standard tanglegram, but is an important property that we leverage to reconstruct widespread events (see next section).

### *Recovering Widespread Events*

The widespread events considered herein are derived from existing divergence events, and therefore existing techniques for the recovery of divergence events may be applied to their recovery. This approach while used by Page (1994b) to infer failure-to-diverge event,s has not been applied to reconcile multiple widespread parasites within a single common framework. To achieve this each widespread event is considered as the divergence event which most closely matches its behaviour. Under this constraint a failure-to-diverge is recovered from an association tree as a codivergence, and a spread is recovered from an association tree as a host switch.

This is possible as both the optimal codivergence and failure-to-diverge events occur at the most recent common ancestor of their children (Johnson et al. 2003), while the optimal host switch and spread events may be recovered using an implementation of the level ancestor problem (Drinkwater and Charleston 2014a). This is possible as each

widespread event mirrors these two divergence events, with the exception that neither include a divergence. This is resolved by creating pseudo-divergence events through the construction of the association trees in line with Siddall and Perkins's, (2003), proposed reconciliation model.

Therefore as both widespread events can be inferred from existing divergence events, we may apply existing solutions to the dated tree reconciliation problem, as a means to recover the optimal divergence events for the set of association trees,  $A$ . This may in-turn be leveraged to infer the optimal set of widespread events for each association tree,  $a_i$ . This is possible as association trees are constructed with a one-to-one mapping, which mitigates the need for duplication events, if host switch events are permitted. This is important as there is no widespread equivalent for a duplication event. Exploiting this imposed property of each association tree, we can reconstruct the map for each association tree where only codivergence, host switch and loss events are permitted; that is running the existing Improved Node Mapping algorithm with a cost vector of  $(F, \infty, S, L)$ , where the costs for failure-to-diverge ( $F$ ) and spread ( $S$ ) replace the costs for codivergence and host switch respectively.

The widespread events are inferred from the recovered mappings by relabelling each codivergence as a failure-to-diverge and each host switch as a spread. This process requires that each divergence event in the resultant dynamic programming table generated by the Improved Node Mapping algorithm may be replaced with its corresponding widespread event. The inferred widespread events are then retained within a dynamic programming table  $d_i$ , which contains all the optimal widespread events for the parasite  $p_i$ . Therefore the result of mapping the complete set of association trees  $A$  into  $H$  gives rise to a set of dynamic programming tables  $\omega = (d_1, \dots d_n)$ , containing all the optimal widespread events for the parasite tree  $P$ .

The ReconcileWidespreadParasite algorithm applied to infer the complete set of

optimal widespread events for a parasite tree  $P$  with respect to its host  $H$  is defined in Figure 4. This process outlines a new approach to reconciling the incongruence caused by widespread parasitism. It integrates a number of existing approaches proposed by Page (1994b), Siddall and Perkins (2003), and Brooks (1991), along with integrating the works of Banks and Paterson (2005), and Johnson et al. (2003) into a single reconciliation methodology within the context of dated trees. This in turn provides the foundations to infer the optimal set of divergence events, which is described in detail in the following section.

### *Recovering Divergence Events*

The recovery of the divergence events using WiSPA is derived from traditional bottom-up (taxa-to-root) dynamic programming approaches applied in the Slicing (Doyon et al. 2010), Edge Mapping (Yodpinyanee et al. 2011), and Improved Node Mapping (Drinkwater and Charleston 2014a) algorithms. Each of these existing approaches incrementally constructs their resultant map using a series of sub-solutions, leading to the recovery of an optimal mapping of the parasite phylogeny into its host.

One such method, the Improved Node Mapping algorithm, is a cubic time solution for the dated tree reconciliation problem. This approach reconciles the incongruence displayed for each parasite node, by reconciling the optimal divergence event based on the set of mapping sites for its children. This requires a nested set of loops so that every mapping site for the left child is compared with every mapping site of the right child.

The WiSPA algorithm unlike Improved Node Mapping considers multiple optimal locations for each parasite node, rather than a single optimal mapping site which has been the premise of all cubic time solutions to this problem (Doyon et al. 2010; Yodpinyanee et al. 2011; Drinkwater and Charleston 2014a). In this more complex case, rather than an optimal mapping site for each pair of children, the optimal mapping may occur at any

location, with the sub-solution defined by the widespread mapping. Initial analysis may suggest that this additional complexity may induce a further set of quadratic comparisons. This, however, can be mitigated by exploiting a number of topological properties of the underlying dynamic programming table and the topology of the resultant map, both of which are explored within this section.

The first point which should be noted is that the dynamic programming table traditionally only retains a single mapping site for the parasite leaves (Drinkwater and Charleston 2015a). It is possible, however, to retain multiple mappings for each parasite leaf node, where in fact there is the ability to retain a mapping site for each parasite node to all locations in the host tree, without increasing the asymptotic complexity of the Improved Node Mapping algorithm. Exploiting this property of the dynamic programming table in handling widespread parasites was introduced as a possibility during the formulation of the original Improved Node Mapping algorithm (Drinkwater and Charleston 2014a). By allowing a set of mappings for each parasite node of this size, allows for the optimal set of mapping sites stored for the root of the association tree  $a_i$ , corresponding to the parasite node  $p_i$  in question, to be retained within the dynamic programming table. This in turn allows for the optimal set of widespread events to be considered within the context of inferring a set of optimal divergence events.

To handle the additional complexity which arises due to handling multiple widespread parasite events, the Improved Node Mapping algorithm has been split such that it considers three possible scenarios, including the case where the left child is treated as a widespread parasite, the right child is treated as a widespread parasite or neither child is treated as a widespread parasite, as can be seen in Figure 5. In the case where the left or right child is treated as a widespread event (lines 17 - 28), the divergence event may be placed at an earlier time period to root of the widespread event (either a failure-to-diverge or a spread event), as long as the relative order of the parasite phylogeny is preserved.

That is while a divergence event may be placed prior to multiple widespread events, it may never be placed at a position prior to one of its descendants. Prior in this context refers to a position closer to the present, as the solutions are constructed in reverse, from the tips to the root. To provide this additional degree of flexibility when reconciling the incongruence between the parasite and its host, requires that all positions within the host be considered as a possible mapping site for each pair of points, adding an additional nested loop (on lines 18-21 and 25-27, in the case where the left or right child are widespread respectively).

Handling widespread parasitism in this fashion results in either a widespread event being the root of a sub-solution, such as a failure-to-diverge event occurring at a time period in the past before any of the divergence events, or a divergence event occurring as the root of a sub-solution. In the latter case this sub-solution from this point onwards is considered as a standard mapping site, in line with previous models, while in the case where the root is a widespread event, its parent too will be required to traverse the complete search space to allocate the optimal divergence event, and therefore an additional layer of computational complexity is added with this approach, discussed in detail in the following section.

The major benefit of this model is that in the case where both the left and right children are widespread parasites, it is possible to abstract away any possible compounding complexity by considering each widespread parasite in series. This reduces the need for an additional increase in the computational complexity of the proposed model, which is achieved by noting that a divergence event may not occur prior to the root of both widespread events, as this would reflect the occurrence of divergence events, and as such one of the two widespread events must be considered as a root, or the divergence event itself may be the root of both lineages. This is in line with the theory considered by Fish, (2013), in the development of the third version of Jane was the first version to consider widespread parasitism.

In the final case (lines 29 - 32) neither the left or right child are rooted by widespread events. In this case the complexity of widespread parasitism is already fully explained within the sub-solution, or the sub-solution does not contain any widespread events. In either case such a sub-solution is processed in-line with the existing Improved Node Mapping algorithm, and no further changes are required to the algorithm presented in Figure 5 to handle this case.

Therefore by reconciling the optimal set of divergence events based on the optimal set of widespread evolutionary events retained within  $\omega$  it is possible to handle multiple widespread evolutionary events and to overcome the limitations identified within the algorithm applied by Jane. In the following section the asymptotic complexity of the algorithm is discussed, where we prove that the additional accuracy provided by the model is achieved by adding only a  $O(n)$  increase in the complexity of the Improved Node Mapping algorithm, resulting in a complexity which is comparable to software tools such as Costscape and Eventscape (Libeskind-Hadas et al. 2014) and significantly faster than Jane 1 (Conow et al. 2010) and the Jungle method (Charleston 1998, 2012), all of which are popular co-evolutionary analysis methods.

### *Complexity Analysis*

The WiSPA algorithm is designed using a series of underlying algorithms to provide the most accurate algorithm for handling widespread parasites. In this section we analyse the associated computational complexity of this approach, and how this compares to existing algorithms applied within the field of coevolutionary analysis of widespread parasites.

For the complexity analysis considered herein we consider the number of nodes in the host tree to be  $2n - 1$ . That is that the host tree contains  $n$  leaves and  $n - 1$  internal nodes. The parasite tree conversely contains  $2m - 1$  nodes, where the parasite tree

contains  $m$  leaves and  $m - 1$  internal nodes. Finally the maximum number of associations for an individual widespread parasite is considered as  $k$  where  $k \leq n$ . That is no single parasite may have more associations then there are unique host leaves to infect.

The WiSPA algorithm is composed of two computationally expensive steps. The first is the processing required to handle the parasites which inhabit more than one host, specifically constructing and solving the association trees (lines 7-12 in Figure 5), and the second step is processing the divergence events, the internal nodes in the parasite tree (lines 14-34 in Figure 5).

Processing the leaves in the parasite tree requires the construction of  $O(m)$  association trees which are of size  $O(k)$ . The association trees are constructed using an application of Lozano et al.'s (2007) homeomorphic subgraph pruning algorithm, which runs in  $O(kn)$  for each of the  $O(m)$  widespread parasites. Therefore the time required to construct the set of association trees,  $A$ , is  $O(kmn)$ . The solutions for each of these association trees are stored with an array of dynamic programming tables, where each table is of size  $O(nk)$ , where the array of dynamic programming tables  $\omega$  contains  $O(m)$  elements. Therefore the space requirement for the step is  $O(kmn)$ . Solving each of the association trees requires  $O(kn^2)$  time, and therefore as  $O(m)$  trees need to be solved the total running time of this step is  $O(kmn^2)$ .

Reconciling the divergence events (lines 14-34 in Figure 5) requires mapping the parasite into the host using the additional information retained within the list of dynamic programming tables,  $\omega$ . As the additional widespread information is retained within  $\omega$  no additional space is required compared to the original dynamic programming table construction defined by Drinkwater and Charleston (2014a), and therefore the space required is  $O(mn)$ . The running time however requires an additional step which involves iterating over all the possible widespread locations of which there may be  $O(k)$  for each mapping site considered, and therefore the running time is extended from  $O(mn^2)$ , as

defined within the original implementation of the Improved Node Mapping algorithm, to  $O(kmn^2)$ .

This time and space complexity is quite significant considering that the complexity of the proposed algorithm grows linearly in regards to the number of additional widespread associations which are added to the tanglegram. That is while the Improved Node Mapping algorithm runs in cubic time when considering only  $O(n)$  associations, our proposed algorithm runs in quartic time when considering  $O(n^2)$  associations. Therefore in the case where only one additional widespread association is added to each parasite, the total running time only increases by a factor of two. This is significant as the number of widespread associations for each parasite will never be of size  $O(n)$  under any realistic biological scenario. For example, if we consider the 15 previously published biological data sets introduced later to validate our model, it may be observed that on average the rate of widespread parasitism is approximately 7%, which compared to the size of the data sets is less than  $\log n$ , which argues that while the worst case running time for the proposed algorithm is quartic, the actual running time in practice is actually more comparable to existing cubic time algorithms.

### *Implementation and Validation*

The algorithm proposed herein is implemented in Java and is available as a platform-independent jar file. The underlying algorithm is integrated into a genetic algorithm, which is designed to run in a multithreaded environment, similar to the design proposed by Conow et al. (2010). The advantage of Conow et al.'s (2010) model is the near-linear speedup possible using multi-core systems.

Jane 4 (Conow et al. 2010) was selected as the algorithm to validate the theoretical model presented herein. Jane is the best candidate to evaluate the performance of WiSPA as both methods are designed to minimise the total cost of all evolutionary events



considered, and that they both leverage an underlying algorithm to solve the dated tree reconciliation problem as a means to inform their metaheuristic framework. CoRe-PA and CoRe-ILP were not considered, as for the size of the data sets considered herein Jane has been shown to outperform both these techniques (Conow et al. 2010; Wieseke et al. 2015).

The evaluation of our new model is broken into two parts. The first considers Jane and WiSPA’s accuracy over 500 synthetic data sets which display varying degrees of widespread parasitism. Then Jane and WiSPA are evaluated over 15 previously published biological systems. In both evaluations two key metrics are considered. The first is the total cost of the reconciliation inferred by each model, and the second is the total number of codivergence events present in the inferred reconciliation. Each of these two values represent the degree of congruence represented by the reconciled map, where the aim for coevolutionary analysis is to infer the minimum cost map with the maximum number of codivergence events (Littlewood 2003). Therefore each model will be validated on how well they conform to this criteria. These two key metrics align with prior analysis of coevolutionary techniques (Page 1994a, 2002; Ronquist 1998; Conow et al. 2010; Wieseke et al. 2015), and are considered the best two signals for recovering a biologically relevant map which most accurately represents the actual coevolutionary interactions.

Along with demonstrating the effectiveness of the generalised model applied within WiSPA, this analysis also aims to infer the significance of the inclusion of the spread evolutionary event. This was achieved by considering three different costs for the evolutionary event spread; a cost of one, which is equal to the cost of a failure-to-diverge event, a cost of two, the same cost as a host switch event the evolutionary event which is most similar to the spread event, and finally the case where a spread event is not permitted (in essence assigned a cost of  $\infty$ ).

Each of these values for the spread event are integrated into the Jungle cost scheme (Ronquist 2003) to provide three unique event cost schemes for this analysis, including

$V = (0, 1, 2, 1, 1, 1)$ ,  $V = (0, 1, 2, 1, 1, 2)$ , and  $V = (0, 1, 2, 1, 1, \infty)$ . When evaluating the performance of Jane using these cost vectors the recovered map will always have the same cost, as varying the cost of spread has no bearing on the cost of the recovered map by Jane.

## DISCUSSION AND ANALYSIS

The analysis performed herein using a combination of synthetic and biological data sets will demonstrate that WiSPA is able to converge on maps with a lower event cost, with a significantly higher number of codivergence events. The significance of this result is that even in the case where spread is not permitted, WiSPA is observed to perform 2% better in practice. This is shown to only improve as spread is permitted, and its associated penalty cost is reduced.

Following this successful result we continue our analysis of WiSPA and Jane by considering the *Primate-Enterobius* biological data sets in further detail. This biological system has long been considered a likely coevolutionary system, however, the inability of prior models to handle the widespread parasitism resulted in no previous model providing a statistically significant coevolutionary hypothesis for the sub-clade considered herein. We demonstrate that while Jane may be unable to provide such a hypothesis, for certain values of spread, WiSPA is able to provide a statistically significant coevolutionary hypothesis for this system.

### *Overall Performance on Synthetic Data*

The synthetic data sets used to evaluate WISPA were previously constructed using the Cophylogeny Generation Model (Core-Gen) (Keller-Schmidt et al. 2011). These coevolutionary histories were constructed using a standard Yule Model, a common synthetic tree generation model applied in phylogenetics (Steel and McKenzie 2001).

Previously Keller-Schmidt et al. (2011) constructed 1000 synthetic data sets, where for this evaluation we have randomly selected 50 of these to provide a baseline for this comparison.

As Core-Gen can only generate coevolutionary systems where each parasite infects a single host, the existing data sets needed to be modified to induce widespread parasitism. From each of the 50 synthetic data sets initially selected, nine additional data sets were created by randomly applying additional widespread associations to the initial data sets. These additional nine new data sets present a varied degree of widespread parasitism, with the aim to model a decreasing rate of host specificity.

For the nine data sets we allowed additional widespread events to be added for each parasite, such that the maximum rate of additional widespread parasitism was 10%, 15%, 20%, 25%, 30%, 35%, 40%, 45%, and 50% of the total available host species for each of the nine data sets respectively. This is applied by selecting each parasite node and allowing the parasite to infect a random number of additional host species between (0 and  $p \times n$ ), where  $p$  is the rate of widespread parasitism for the specified synthetic system and  $n$  is the number of host taxa. It should be noted that this model is a crude representation of widespread parasitism in nature, however, it provides a robust set of synthetic data sets to compare Jane and WiSPA over varying degrees of widespread parasitism.

This technique is also advantageous as it provides a baseline of the number of codivergence events present in the original tanglegram, which can be compared to the recovered number of codivergence events of each technique as the rate of widespread parasitism increases. Therefore both the rate at which the total event cost increases and the total number of codivergence events decreases, may be tracked for the models considered within this analysis.

This is captured in Figure 6 where the total event cost (left) and total number of codivergence events (right) are recorded for the ten data sets (including the baseline) for the four models considered. These two plots provide the best insight to date in regards to

the benefits of the spread event, particularly in increasing the total number of codivergence events compared to using failure-to-diverge exclusively.

In the case where spread is set to one a reduction of more than 50% is achieved in the parsimony score, with this reduction only increasing as the rate of widespread parasitism increases. This reduction is complimented by a nine fold increase in the number of codivergence events. A similar trend is observed where spread is set to two. Here a reduction of more than 35% is achieved in the parsimony score, with this reduction only increasing as the rate of widespread parasitism increases. The reduction in this case is complimented by an eight fold increase in the number of codivergence events.

In both cases where spread is permitted a significant improvement in the congruence of the reconciled maps is achieved. In the case where spread is not permitted such a drastic improvement is not observed although there is a 1% decrease in the total parsimony cost compared to Jane with an increase of 2% in the number of codivergence events. While nowhere near as impressive as the case where spread is permitted, this improvement is important as while it represents our model in the worst case it still shows it outperforms Jane and only improves as the cost of spread is decreased.

### *Overall Performance on Real Data*

The performance over the synthetic data set demonstrates the value of the generalised model applied within WiSPA along with the advantage of applying the spread event for the analysis of systems presenting widespread parasitism. As noted, however, the model applied to generate the synthetic data sets does not provide the best representation of widespread parasitism within a biological system, although it is the first synthetic model which attempts to model this phenomenon. Therefore our analysis also compares the performance of Jane and WiSPA over biological data sets.

For this analysis 15 biological data sets were selected to compare WiSPA with the latest version of Jane. These biological systems considered 10 biological phenomena, including but not limited to parasitism (Hafner and Nadler 1988), plant–insect interactions (Gómez-Acevedo et al. 2010), coevolutionary dynamics between a virus and its host (Jackson and Charleston 2004), mutualism (McLeish and Van Noort 2012), parasitoidism (Murray et al. 2013), and plant–fungal coevolution (Refrégier et al. 2008). The complete list of biological systems included in this analysis and the coevolutionary interrelationships each system expresses has been listed in Table 1. These data sets were selected to evaluate whether spread can assist in recovering more parsimonious reconstructions, along with evaluating the newly proposed model for reconciling widespread parasitism. This analysis along with evaluating the cost of each reconciliation, also considers the number of codivergence events recovered as another means of evaluating the inferred congruence found by each technique considered herein.

It should be noted that while 15 data sets does not compare to the 500 data sets considered in the previous section, this selection of biological data represents the largest collection of coevolutionary systems displaying widespread parasitism assembled to date. While larger collections exist for the case where parasite only parasites infect a single host such as the 102 data sets catalogued by Drinkwater and Charleston (2014b) this is the first and therefore largest selection of widespread coevolutionary systems aggregated to date.

The results of this comparison are displayed in Table 2 and show a significant improvement in both the reconciliation’s cost, and the total number of codivergence events when including the spread event in the reconstruction of the parasites’ evolutionary history with respect to their host. In all cases where spread was assigned a cost of one, the newly proposed algorithm found a solution that was at least as parsimonious as Jane, with the majority of cases inferring a solution which was significantly more parsimonious and with a higher number of codivergence events. Over the 15 data sets there was an observed

reduction of 35% in the event cost, and an increase of 21% in the number of codivergence events.

A similar trend was observed in the case where spread was assigned a cost of 2. Here in all cases WiSPA was able to find a solution which was at least as parsimonious in terms of event cost, with a number of cases where WiSPA was able to infer a reconciliation which was significantly more parsimonious and with a higher number of codivergence events. Over the 15 data sets there was an observed reduction of 22% in the event cost and an increase of 18% in the number of codivergence events.

This demonstrates the value of the spread event for coevolutionary analysis, where a significant reduction in the total parsimony cost may be achieved using the spread event, with the largest single reduction providing a 55% decrease in the total event cost. These results match the benefits observed over the synthetic data sets, providing further evidence of the value of adopting the spread event for widespread analysis.

In the case where spread was not permitted, WiSPA was still able to outperform Jane, although the improvement was not as pronounced. Overall there was a 2% reduction in the total cost of the 15 maps with no difference in the total number of codivergence events inferred across the 15 systems. This is still a significant result as this is the worst case performance of the newly proposed model for reconciling widespread parasitism, and even then we are able to present an improvement of 2%. While in the case where spread is not permitted many systems perform as well as Jane, there is one particular system which displays a significant improvement. The ant-wasp parasitoid coevolutionary system's cost is reduced by 25% by allowing a divergence event to occur prior to multiple widespread events. Such a significant reduction is the difference between a map which is only cheaper than 61.19% of random solutions as in the case of Jane compared to a map which is cheaper than 94.96% of random solutions as in the case of WiSPA. These results are based on the randomisation test undertaken using Jane, using 10000 random instances.

While in all cases WiSPA was able to recover a map which was equal to or less than that which was recovered by Jane it should be noted that there was a case where Jane was able to outperform WiSPA in terms of inferring a map where the total number of codivergence events was higher. In the RNA Virus example, which has been marked as bold in Table 2, it can be seen that Jane’s best reconciliation contains 5 codivergence events while WiSPA only infers a map with 4. In both cases the recovered map has a cost of 15 and as such current significance testing considers these two model equivalent. If subjected to the same randomisation test considered for the ant–wasp parasitoid coevolutionary systems neither map is considered significant, ( $p = 0.14$ ).

The results over the biological data sets provide a compelling case for the adoption of the newly proposed model. Here a reduction in the parsimony score of 22% is achieved even where spread is set to a cost of two, reducing a further 14% if spread is assigned a cost of one. In the following section we demonstrate the significance of this improvement by comparing Jane’s reconciliation with those of our model along over a long studied biological system of public health significance.

### *Spread provides stronger evidence for Primate–Enterobius Coevolution*

*Primate* and *Enterobius* (Pinworms) have long been considered as a possible coevolutionary system (Cameron 1929; Sandosham 1950; Sorci et al. 1997; Hugot 1999). This hypothesis is due to the high degree of congruence which has been observed between these two phylogenetic trees (Hugot 1999). Brooks and Glen (1982) identified that while the observed congruence within this system strongly supported coevolution that there remains a subset of the *Primate* and *Enterobius* tanglegram which did not appear to provide evidence of coevolution. In particular it was noted that the species *E. vermicularies* infection of both *Hylobatidae* (Gibbon) and *Homo sapien* (Human) could not be explained by traditional coevolutionary models.

This failure by cladistic models was due to the inability of coevolutionary analysis to reconcile widespread parasitism. This was later rectified as part of SBPA which Brooks and McLennan 2003 applied to this system although once again no specific modelling for the relationships between the species *E. vermicularies* which inhabits both humans and gibbons was provided. This unexplained sub-clade has also been considered by cophylogeny models as well as cladistic approaches, with Ronquist (1997) proposing that this inconsistency was due to a recent host switch event from gibbons to humans. One weakness with this hypothesis, however, is that it assumes that *E. vermicularies* has diverged during the infection of Humans which current evidence does not support (Brooks and Glen 1982). As a result, a complete hypothesis which reconciles the observed data within this potential coevolutionary system and in particular this sub-clade has remained unanswered.

While initial coevolutionary analysis assumed widespread parasitism cannot occur (Poulin 2011) in a coevolutionary context, this has gradually become more accepted as potentially occurring depending on the nature of the coevolutionary system considered. Laboratory experiments have shown that *Enterobius* is a species which displays a low host specificity, with results as early as Sandosham (1950) noting that a number of species of *Enterobius* infected phylogenetically distant primates held within captivity and which would not associate with one another in the wild due to vast geographical diversity. From this evidence it does not seem infeasible that *E. vermicularies* may also be able to infect multiple host species.

The infection of humans has a higher probability due to humans no longer being bound by their biogeographical environment. This hypothesis agrees with existing coevolutionary analysis focusing on tapeworms, which have been shown to have a low host specificity whereby species were able to infect humans during their dispersal from Africa 2.5 million years ago (Hoberg et al. 2001).

Therefore, we attempt to provide a coevolutionary explanation applying widespread



parasitism to this sub-clade using both the methodologies applied in Jane and WiSPA. We firstly evaluate the two recovered maps from Jane and WiSPA and discuss their inferred set of biological events and their implications. These maps are then evaluated statistically to evaluate if either method rejects the null hypothesis that these two phylogenetic trees are independent from one another. For this analysis we apply the Jungle cost scheme (Ronquist 2003) including spread with a cost of both one and two.

To provide a fair statistical analysis we generate all feasible widespread systems which include one additional widespread association between the parasite and its host. In total there are 10000 systems where the host and parasite phylogenies are fixed and the associations are randomised. By generating a single instance of all possible maps, we guarantee that no bias is introduced using different randomisation techniques for each model.

These models may be generated by computing all possible association pairs of which there are  $5^4$  and multiplying this by the total number of unique additional associations that may be applied, which is  $(5 - 1) \times 4$ . The minus one is due to the inability to apply more than one association between a single parasite and a single host. Therefore the total number of unique systems that may be generated for the *Primate* and *Enterobius* (Pinworms) system presented in Figure 7 is 10000 ( $5^4 \times (5 - 1) \times 4$ ).

Jane's recovered map for the Jungle cost scheme is visualised in Figure 8 (left). This map consists of two codivergences, one host switch, three loss events and one failure-to-diverge which has a resultant cost of six. This map hypothesises that the species of pinworm which now infects gibbon and human had the potential of infecting all species which humans share a more recent ancestor with than with gibbons but that this species failed to do so on all occasions. This seems highly improbable as gibbon and human's most recent common ancestor is estimated to have lived over 14 million years ago (Carbone et al. 2014).

Further, while the initial parasite phylogeny appears to have a high degree of congruence with its host, this apparent congruence is not well represented in the recovered map. If we compare Jane’s recovered map to the inferred cost over the 10000 unique instances, we observe (in Figure 9 (left)) that it is relatively high compared to what would be expected simply by chance. If we evaluate its cost using the Wilson score interval we converge on a confidence interval of (0.760, 0.795). As such the reconciliation argues that the evaluation history of these two sets of species are independent.

WiSPA’s recovered map for the cost scheme  $V = (0, 1, 1, 2, 1, 1)$  and  $V = (0, 1, 1, 2, 1, 2)$  is visualised in Figure 8 (right). In both cases the same map was inferred where the only difference was that the recovered spread event costs more in the latter case. This map consists of three codivergences, one loss event and one spread which has a resultant cost of two or three respectively. This map provides the hypothesis that this system has been coevolving throughout its evolutionary history with all divergence events indicative of coevolution. This widespread event for *E. vermicularies* in this map is explained using a recent spread event from gibbon to human. As previously discussed spread requires that the hosts are biologically collocated at the time spread occurs. This collocation can be explained in this case as humans are no longer geographically bound and therefore spread’s potential is significantly higher than for other geographically bound primates. The loss event in this reconstruction can be explained by integrating citeauthorronquist1997phylogenetic’s (1997) prior hypothesis. In particular he noted the introduction of *E. vermicularies* into humans may have caused an extinction of a species of *Enterobius* with a common ancestor of *E. anthropopithecii*.

If we compare WiSPA’s recovered map to the same 10000 unique instances that were used to evaluate Jane’s map, it can be seen that there is strong evidence for coevolution in this case, as seen in Figure 9 (center and right). Using the Wilson score interval we converge on a confidence interval for the case where spread is a cost of one and two of

(0.028, 0.045) and (0.044, 0.069) respectively. In both cases this presents a strong indication of coevolution, considering the size of the tanglegram where even a perfectly congruent tanglegram with four internal nodes in the host and parasite tree (an additional node in the parasite compared to this example) only offers a confidence value of (0.011, 0.023).

Due to these results we argue that the inconsistent sub-clade from the *Primate* / *Enterobius* tanglegram can be explained using widespread parasitism when applying the spread event. In particular we note that the algorithm WiSPA is the only method that is able to recover a widespread solution to this instance and provide a statistically significant signal for coevolution for this evolutionary system.

## CONCLUSION

This work presents a new model for reconciling the incongruence that may arise between a pair of phylogenetic trees where parasites are permitted to inhabit more than one host. While this permutation of the cophylogeny reconstruction problem has often been considered to be computationally complex, we provide a polynomial solution in the case where there exists timing information for the host phylogeny. In the case where such timing information is unavailable, we provide a metaheuristic framework which applies our underlying algorithm, which is shown to be the most accurate model for widespread parasitism produced to date.

The accuracy improvement present within our proposed model (WiSPA) is due to its inclusion of an additional widespread evolutionary event, which we refer to as spread, along with it providing a more generalised framework for inferring the optimal set of widespread events. The additional widespread evolutionary event applied herein is derived from a number of previous coevolutionary models, along with observed parasitic behaviour in nature and the laboratory, where the inclusion of the spread event alone has been shown to provide an accuracy improvement of over 55%.

The accuracy improvement comes at a cost however, where our model is shown to be an order of magnitude slower than the current state of the art algorithm applied in Jane. While this algorithm is more computationally expensive than algorithms applied in the latest version of Jane (Libeskind-Hadas 2015) and CoRe-PA (Merkle et al. 2010), our algorithm is still far superior to Jane 1 (Conow et al. 2010) and the Jungle model applied in TreeMap, which have both been applied to successfully analyse a number of coevolutionary systems, and is also asymptotically more efficient than the tools within the Xscape framework (Libeskind-Hadas et al. 2014), proving that our model is capable of analysing biological data sets.

Finally we applied WiSPA to the well-studied sub-clade of the coevolutionary system of *Primate* and *Enterobius*. Since this sub-clade was identified by Brooks and Glen (1982) no satisfactory explanation reconciliation of this sub-clade has been derived. We have shown that while this has eluded prior models, WiSPA is able to provide a statistically significant hypothesis for this sub-clade which complements the existing theory of *Primate* / *Enterobius* coevolution, and also provides a plausible biological model consistent with broader understanding of primate–parasite coevolution.

This result coupled with the results when comparing WiSPA and Jane over the synthetic and biological data sets considered herein demonstrates the value of our proposed model. Not only does WiSPA provide the flexibility of providing an additional evolutionary event to explain the incongruence caused by widespread taxa but this model also provides further flexibility in reconciling the conflict that may arise when dealing with the order of widespread and divergence events. As such we argue for the adoption of this new model to provide additional insights into the complex problem of reconciling the coevolutionary associations of widespread taxa.

## References

- Banks, J. and A. Paterson. 2005. Multi-Host Parasite Species in Cophylogenetic Studies. *International Journal for Parasitology* 35:741–746.
- Bansal, M. S., E. J. Alm, and M. Kellis. 2012. Efficient algorithms for the reconciliation problem with gene duplication, horizontal transfer and loss. *Bioinformatics* 28:i283–i291.
- Baudet, C., B. Donati, B. Sinimeri, P. Crescenzi, C. Gautier, C. Matias, and M.-F. Sagot. 2015. Cophylogeny Reconstruction via an Approximate Bayesian Computation. *Systematic Biology* 64:416–431.
- Brooks, D. R. 1991. *Phylogeny, ecology, and behavior: a research program in comparative biology*. University of Chicago Press.
- Brooks, D. R. and D. R. Glen. 1982. Pinworms and primates: a case study in coevolution. *Proceedings of the Helminthological Society of Washington* 49:76–85.
- Brooks, D. R. and D. A. McLennan. 2003. Extending phylogenetic studies of coevolution: secondary Brooks parsimony analysis, parasites, and the Great Apes. *Cladistics* 19:104–119.
- Cameron, T. 1929. The species of *Enterobius* Leach, in primates. *Journal of Helminthology* 7:161–182.
- Carbone, L., R. A. Harris, S. Gnerre, K. R. Veeramah, B. Lorente-Galdos, J. Huddleston, T. J. Meyer, J. Herrero, C. Roos, B. Aken, et al. 2014. Gibbon genome and the fast karyotype evolution of small apes. *Nature* 513:195–201.

- Ceccarelli, F. and R. Crozier. 2007. Dynamics of the evolution of Batesian mimicry: molecular phylogenetic analysis of ant-mimicking *Myrmarachne* (Araneae: Salticidae) species and their ant models. *Journal of Evolutionary Biology* 20:286–295.
- Charleston, M. 1998. Jungles: A new solution to the Host/Parasite Phylogeny Reconciliation Problem. *Mathematical Biosciences* 149:191–223.
- Charleston, M. 2012. Download TreeMap 3 [here](#).
- Charleston, M. and A. Galvani. 2006. A cophylogenetic perspective on host-pathogen evolution. *DIMACS Series in Discrete Mathematics and Theoretical Computer Science* 71:145.
- Charleston, M. and R. Libeskind-Hadas. 2014. Event-Based Cophylogenetic Comparative Analysis. Pages 465–480 *in* *Modern Phylogenetic Comparative Methods and Their Application in Evolutionary Biology*. Springer.
- Charleston, M. and D. Robertson. 2002. Preferential host switching by primate lentiviruses can account for phylogenetic similarity with the primate phylogeny. *Systematic Biology* 51:528–535.
- Charleston, M. A. 2002. Principles of cophylogenetic maps. Pages 122–147 *in* *Biological Evolution and Statistical Physics*. Springer.
- Charleston, M. A. 2003. Recent results in cophylogeny mapping. *Advances in parasitology* 54:303–330.
- Charleston, M. A. and S. L. Perkins. 2006. Traversing the Tangle: Algorithms and Applications for Cophylogenetic Studies. *Journal of Biomedical Informatics* 39:62–71.
- Clayton, D. H., S. E. Bush, and K. P. Johnson. 2004. Ecology of congruence: past meets present. *Systematic Biology* 53:165–173.

- Conow, C., D. Fielder, Y. Ovadia, and R. Libeskind-Hadas. 2010. Jane: a new tool for the Cophylogeny Reconstruction Problem. *Algorithms for Molecular Biology* 5:16.
- Cruaud, A., N. Rønsted, B. Chantarasuwan, L. S. Chou, W. L. Clement, A. Couloux, B. Cousins, G. Genson, R. D. Harrison, P. E. Hanson, et al. 2012. An extreme case of plant–insect codiversification: figs and fig-pollinating wasps. *Systematic Biology* 61:1029–1047.
- Cuthill, J. H. and M. Charleston. 2012. Phylogenetic Codivergence Supports Coevolution of Mimetic *Heliconius* Butterflies. *PloS One* 7:e36464.
- Cuthill, J. H. and M. A. Charleston. 2013. A simple model explains the dynamics of preferential host switching among mammal RNA viruses. *Evolution* 67:980–990.
- Doyon, J.-P., V. Ranwez, V. Daubin, and V. Berry. 2011. Models, Algorithms and Programs for Phylogeny Reconciliation. *Briefings in Bioinformatics* 12:392–400.
- Doyon, J.-P., C. Scornavacca, K. Y. Gorbunov, G. J. Szöllősi, V. Ranwez, and V. Berry. 2010. An Efficient Algorithm for Gene / Species Trees Parsimonious Reconciliation with Losses, Duplications and Transfers. Pages 93–108 *in Comparative Genomics*. Springer.
- Drinkwater, B. and M. Charleston. 2016. RASCAL: A randomised approach for coevolutionary analysis. *Journal of Computational Biology* 23:218–227.
- Drinkwater, B. and M. A. Charleston. 2014a. An Improved Node Mapping Algorithm for the Cophylogeny Reconstruction Problem. *Coevolution* 2:1–17.
- Drinkwater, B. and M. A. Charleston. 2014b. Introducing TreeCollapse: A novel greedy algorithm to solve the Cophylogeny Reconstruction Problem. *BMC Bioinformatics* 15:S14.

- Drinkwater, B. and M. A. Charleston. 2015a. A Sub-quadratic Time and Space Complexity Solution for the Dated Tree Reconciliation Problem for Select Tree Topologies. Pages 93–107 *in* Algorithms in Bioinformatics. Springer.
- Drinkwater, B. and M. A. Charleston. 2015b. A time and space complexity reduction for coevolutionary analysis of trees generated under both a Yule and Uniform model. Computational biology and chemistry 57:61–71.
- Escudero, M. 2015. Phylogenetic congruence of parasitic smut fungi (*Anthracoidea*, *Anthracoideaceae*) and their host plants (*Carex*, *Cyperaceae*): Cospeciation or host-shift speciation? American Journal of Botany 102:1108–1114.
- Fahrenholz, H. 1913. Ectoparasiten und abstammungslehre. Zoologischer Anzeiger 41:371–374.
- Fish, B. 2013. The Cophylogeny Reconstruction Problem. Pomona College.
- Gómez-Acevedo, S., L. Rico-Arce, A. Delgado-Salinas, S. Magallón, and L. E. Eguiarte. 2010. Neotropical mutualism between Acacia and Pseudomyrmex: phylogeny and divergence times. Molecular Phylogenetics and Evolution 56:393–408.
- Hafner, M. S. and S. A. Nadler. 1988. Phylogenetic trees support the coevolution of parasites and their hosts. Nature .
- Hendricks, S. A., M. E. Flannery, and G. S. Spicer. 2013. Cophylogeny of quill mites from the genus *Syringophilopsis* (Acari: *Syringophilidae*) and their North American passerine hosts. The Journal of Parasitology 99:827–834.
- Hoberg, E. P., N. L. Alkire, A. Queiroz, and A. Jones. 2001. Out of Africa: origins of the Taenia tapeworms in humans. Proceedings of The Royal Society of London B: Biological Sciences 268:781–787.



- Hugot, J. 1999. Primates and their pinworm parasites: the Cameron hypothesis revisited. *Systematic Biology* 48:523–546.
- Jackson, A. P. and M. A. Charleston. 2004. A cophylogenetic perspective of RNA–virus evolution. *Molecular Biology and Evolution* 21:45–57.
- Jaenike, J. and I. Dombeck. 1998. General-purpose genotypes for host species utilization in a nematode parasite of *Drosophila*. *Evolution Pages* 832–840.
- Johnson, K. P., R. J. Adams, R. D. Page, and D. H. Clayton. 2003. When do parasites fail to speciate in response to host speciation? *Systematic Biology* 52:37–47.
- Keller-Schmidt, S., N. Wieseke, K. Klemm, and M. Middendorf. 2011. Evaluation of host parasite reconciliation methods using a new approach for cophylogeny generation. Tech. rep. Working paper from Bioinformatics Leipzig. Available from [http://www. bioinf. uni-leipzig. de/working/11-013](http://www.bioinf.uni-leipzig.de/working/11-013).
- Kellner, K., H. Fernández-Marín, H. Ishak, R. Sen, T. Linksvayer, and U. Mueller. 2013. Co-evolutionary patterns and diversification of ant–fungus associations in the asexual fungus-farming ant *Mycocepurus smithii* in Panama. *Journal of Evolutionary Biology* 26:1353–1362.
- Kim, K. C. et al. 1985. *Coevolution of parasitic arthropods and mammals*. John Wiley and Sons.
- Legendre, P., Y. Desdevises, and E. Bazin. 2002. A statistical test for host–parasite coevolution. *Systematic Biology* 51:217–234.
- Libeskind-Hadas, R. 2015. Who is Jane?

Libeskind-Hadas, R. and M. Charleston. 2009. On the Computational Complexity of the Reticulate Cophylogeny Reconstruction Problem. *Journal of Computational Biology* 16:105–117.

Libeskind-Hadas, R., Y.-C. Wu, M. S. Bansal, and M. Kellis. 2014. Pareto-optimal phylogenetic tree reconciliation. *Bioinformatics* 30:i87–i95.

Littlewood, T. 2003. *The Evolution of Parasitism—A Phylogenetic Perspective* vol. 54. Academic Press.

Lockyer, A., P. Olson, P. Ostergaard, D. Rollinson, D. Johnston, S. Attwood, V. Southgate, P. Horak, S. Snyder, T. Le, et al. 2003. The phylogeny of the Schistosomatidae based on three genes with emphasis on the interrelationships of *Schistosoma* Weinland, 1858. *Parasitology* 126:203–224.

Lozano, A., R. Y. Pinter, O. Rokhlenko, G. Valiente, and M. Ziv-Ukelson. 2007. Seeded tree alignment and planar tanglegram layout. Pages 98–110 *in* *Algorithms in Bioinformatics*. Springer.

Martínez-Aquino, A., F. S. Ceccarelli, L. E. Eguiarte, E. Vázquez-Domínguez, and G. P.-P. de León. 2014. Do the historical biogeography and evolutionary history of the digenean *margotrema* spp. across central mexico mirror those of their freshwater fish hosts (goodeinae)? *PLoS One* 9:e101700.

McLeish, M. J. and S. Van Noort. 2012. Codivergence and multiple Host Species use by Fig Wasp Populations of the Ficus Pollination Mutualism. *BMC Evolutionary Biology* 12:1.

Mendlova, M., Y. Desdevises, K. Civaňová, A. Pariselle, and A. Šimková. 2012. Monogeneans of West African cichlid fish: evolution and cophylogenetic interactions. *PLoS One* 7:e37268.

- Merkle, D. and M. Middendorf. 2005. Reconstruction of the cophylogenetic history of related phylogenetic trees with divergence timing information. *Theory in Biosciences* 123:277–299.
- Merkle, D., M. Middendorf, and N. Wieseke. 2010. A parameter-adaptive dynamic programming approach for inferring cophylogenies. *BMC Bioinformatics* 11:S60.
- Mu, J., D. A. Joy, J. Duan, Y. Huang, J. Carlton, J. Walker, J. Barnwell, P. Beerli, M. Charleston, O. Pybus, et al. 2005. Host switch leads to emergence of plasmodium vivax malaria in humans. *Molecular Biology and Evolution* 22:1686–1693.
- Murray, E. A., A. E. Carmichael, and J. M. Heraty. 2013. Ancient host shifts followed by host conservatism in a group of ant parasitoids. *Proceedings of the Royal Society of London B: Biological Sciences* 280:20130495.
- Nosil, P. and A. Mooers. 2005. Testing hypotheses about ecological specialization using phylogenetic trees. *Evolution* 59:2256–2263.
- Ovadia, Y., D. Fielder, C. Conow, and R. Libeskind-Hadas. 2011. The Cophylogeny Reconstruction Problem is NP-Complete. *Journal of Computational Biology* 18:59–65.
- Page, R. D., R. H. Cruickshank, M. Dickens, R. W. Furness, M. Kennedy, R. L. Palma, and V. S. Smith. 2004. Phylogeny of *Philoceanus complex* seabird lice (Phthiraptera: Ischnocera) inferred from mitochondrial dna sequences. *Molecular Phylogenetics and Evolution* 30:633–652.
- Page, R. D. M. 1994a. Maps Between Trees and Cladistic Analysis of Historical Associations Among Genes, organisms, and areas. *Systematic Biology* 43:58–77.
- Page, R. D. M. 1994b. Parallel Phylogenies: Reconstructing the History of Host-Parasite Assemblages. *Cladistics* 10:155–173.

- Page, R. D. M. 2002. Tangled Trees: Phylogeny, Cospeciation, and Coevolution. University of Chicago Press, Chicago.
- Page, R. D. M. and M. A. Charleston. 1997. From gene to organismal phylogeny: reconciled trees and the gene tree/species tree problem. *Molecular Phylogenetics and Evolution* 7:231–240.
- Paterson, A. and R. Poulin. 1999. Have Chondracanthid Copepods co-specified with their Teleost hosts? *Systematic Parasitology* 44:79–85.
- Paterson, A. M. and J. Banks. 2001. Analytical approaches to measuring cospeciation of host and parasites: through a glass, darkly. *International Journal for Parasitology* 31:1012–1022.
- Paterson, A. M., R. L. Palma, and R. D. Gray. 2003. Drowning on arrival, missing the boat, and X-events: How likely are sorting events. *Tangled Trees: Phylogeny, Cospeciation, and Coevolution* Pages 287–309.
- Peterson, A. T., J. T. Bauer, and J. N. Mills. 2004. Ecologic and geographic distribution of filovirus disease .
- Poulin, R. 2011. *Evolutionary Ecology of Parasites*. Princeton University Press.
- Refrégier, G., M. Le Gac, F. Jabbour, A. Widmer, J. A. Shykoff, R. Yockteng, M. E. Hood, and T. Giraud. 2008. Cophylogeny of the anther smut fungi and their caryophyllaceous hosts: prevalence of host shifts and importance of delimiting parasite species for inferring cospeciation. *BMC Evolutionary Biology* 8:100.
- Rivera-Parra, J. L., I. I. Levin, K. P. Johnson, and P. G. Parker. 2015. Lineage sorting in multihost parasites: *Eidmanniella albescens* and *fregatiella aurifasciata* on seabirds from the Galapagos Islands. *Ecology and Evolution* .

- Ronquist, F. 1995. Reconstructing the history of host-parasite associations using generalised parsimony. *Cladistics* 11:73–89.
- Ronquist, F. 1997. Phylogenetic approaches in coevolution and biogeography. *Zoologica scripta* 26:313–322.
- Ronquist, F. 1998. Three-Dimensional Cost-Matrix Optimization and Maximum Cospeciation. *Cladistics* 14:167–172.
- Ronquist, F. 2003. Parsimony analysis of coevolving species associations. *Tangled Trees: Phylogeny, Cospeciation and Coevolution* Pages 22–64.
- Sandosham, A. 1950. On *Enterobius vermicularis* (Linnaeus, 1758) and Some Related Species from Primates and Rodent. *Journal of Helminthology* 24:171–204.
- Siddall, M. E. 1997. The AIDS pandemic is new, but is HIV not new? *Cladistics* 13:267–273.
- Siddall, M. E. and S. L. Perkins. 2003. Brooks Parsimony Analysis: a valiant failure. *Cladistics* 19:554–564.
- Sorci, G., S. Morand, and J.-P. Hugot. 1997. Host–parasite coevolution: comparative evidence for covariation of life history traits in primates and oxyurid parasites. *Proceedings of the Royal Society of London B: Biological Sciences* 264:285–289.
- Steel, M. and A. McKenzie. 2001. Properties of phylogenetic trees generated by Yule-type speciation models. *Mathematical Biosciences* 170:91–112.
- Stireman, J. 2005. The evolution of generalization? Parasitoid flies and the perils of inferring host range evolution from phylogenies. *Journal of Evolutionary Biology* 18:325–336.

- Toit, N., B. Vuuren, S. Matthee, and C. Matthee. 2013. Biogeography and host-related factors trump parasite life history: limited congruence among the genetic structures of specific ectoparasitic lice and their rodent hosts. *Molecular Ecology* 22:5185–5204.
- Tuller, T., H. Birin, U. Gophna, M. Kupiec, and E. Ruppin. 2010. Reconstructing ancestral gene content by coevolution. *Genome Research* 20:122–132.
- Viale, E., I. Martinez-Sañudo, J. Brown, M. Simonato, V. Girolami, A. Squartini, A. Bressan, M. Faccoli, and L. Mazzon. 2015. Pattern of association between endemic Hawaiian fruit flies (Diptera, Tephritidae) and their symbiotic bacteria: Evidence of cospeciation events and proposal of “*Candidatus Stammerula trupaneae*”. *Molecular Phylogenetics and Evolution* 90:67–79.
- Weckstein, J. 2004. Biogeography explains Cophylogenetic patterns in Toucan Chewing lice. *Systematic Biology* 53:154–164.
- Wieseke, N., T. Hartmann, M. Bernt, and M. Middendorf. 2015. Cophylogenetic Reconciliation with ILP. *IEEE/ACM Transactions on Computational Biology and Bioinformatics* .
- Yodpinyanee, A., B. Cousins, J. Peebles, T. Schramm, and R. Libeskind-Hadas. 2011. Faster Dynamic Programming Algorithms for the Cophylogeny Reconstruction Problem. HMC CS Technical Report .

Table 1: Biological systems considered in this analysis and the type of coevolutionary inter-relationship expressed within said system.

Coevolutionary system	Type of coevolutionary interrelationship expressed
<i>Acacia</i> / <i>Pseudomyrmex</i> (Gómez-Acevedo et al. 2010)	Plant–Insect Mutualism
<i>Aves</i> / <i>Syringophilopsis</i> (Hendricks et al. 2013)	Bird–Mites Parasitism
<i>Carex</i> / <i>Anthracoidea</i> (Escudero 2015)	Plant–Fungi Parasitism
<i>Caryophyllaceae</i> / <i>Microbotryum</i> (Refrégier et al. 2008)	Plant–Fungi Mutualism
<i>Cichlidae</i> / <i>Platyhelminthes</i> (Mendlova et al. 2012)	Fish–Flatworm Parasitism
<i>Formicidae</i> / <i>Eucharitidae</i> (Murray et al. 2013)	Ant–Wasp Parasitoidism
<i>Goodeinae</i> / <i>Margotrema</i> (Martínez-Aquino et al. 2014)	Fish–Flatworm Parasitism
<i>Ficus</i> / <i>Agaonidae</i> (McLeish and Van Noort 2012)	Plant–Insect Mutualism
<i>Gastropoda</i> / <i>Schistosoma</i> (Lockyer et al. 2003)	Snails–Flatworm Parasitism
<i>Geomyidae</i> / <i>Mallophaga</i> (Hafner and Nadler 1988)	Rodent–Lice Parasitism
<i>Mycocepurus smithii</i> / <i>Fungi</i> (Kellner et al. 2013)	Ant–Fungal Mutualism
<i>Ramphastidae</i> / <i>Mallophaga</i> (Weckstein 2004)	Bird–Lice Parasitism
<i>Sigmodontinae</i> / <i>Arenaviridae</i> (Jackson and Charleston 2004)	Rodent–Viral Coevolution
<i>Teleostei</i> / <i>Copepods</i> (Paterson and Poulin 1999)	Fish–Crustacean Parasitism
<i>Tephritidae</i> / <i>Bacteria</i> (Viale et al. 2015)	Fly–Bacteria Symbiosis

Table 2: WiSPA’s performance against Jane 4 over fifteen biological test cases. WiSPA has been run with three different costs associated for spread. Spread was set to a cost of 1, 2 and where spread was not permitted in the reconstruction.

Coevolutionary system	Recovered event cost and (# of codivergence events)			
	Jane	Spread = 1	Spread = 2	No Spread
<i>Acacia</i> / <i>Pseudomyrmex</i>	67 (0)	28 (2)	43 (2)	65 (0)
<i>Aves</i> / <i>Syringophilopsis</i>	17 (9)	17 (9)	17 (9)	17 (9)
<i>Carex</i> / <i>Anthracoidea</i>	73 (9)	59 (9)	65(10)	73 (9)
<i>Caryophyllaceae</i> / <i>Microbotryum</i>	33 (3)	26 (5)	30 (3)	33 (3)
<i>Cichlidae</i> / <i>Platyhelminthes</i>	40 (7)	34 (9)	39 (7)	39 (7)
<i>Formicidae</i> / <i>Eucharitidae</i>	12 (0)	8 (1)	9 (1)	9 (1)
<i>Goodeinae</i> / <i>Margotrema</i>	36 (2)	21 (4)	25 (4)	33 (2)
<i>Ficus</i> / <i>Agaonidae</i>	10 (3)	8 (4)	9 (4)	10 (3)
<i>Gastropoda</i> / <i>Schistosoma</i>	122 (1)	54 (3)	77 (2)	120 (1)
<i>Geomyidae</i> / <i>Mallophaga</i>	9 (6)	8 (6)	9 (6)	9 (6)
<i>Mycocephalus smithii</i> / <i>Fungi</i>	42 (1)	21 (2)	28 (3)	41 (1)
<i>Ramphastidae</i> / <i>Mallophaga</i>	17 (2)	12 (2)	14 (3)	17 (2)
<b><i>Sigmodontinae</i> / <i>Arenaviridae</i></b>	<b>15 (5)</b>	<b>15 (4)</b>	<b>15 (4)</b>	<b>15 (4)</b>
<i>Teleostei</i> / <i>Copepods</i>	4 (1)	2 (2)	3 (2)	4 (1)
<i>Tephritidae</i> / <i>Bacteria</i>	29 (12)	29 (12)	29 (12)	29 (12)
Total	526 (61)	342 (74)	412 (72)	512 (61)



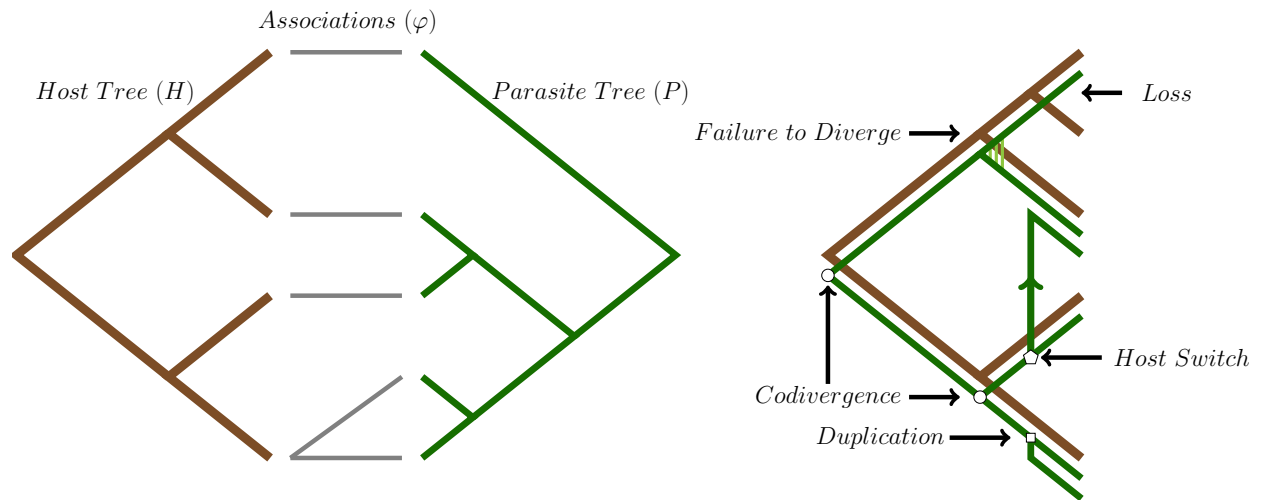


Figure 1: A tanglegram (left) and one of its optimal maps (right). What is unique about this possible map,  $\Phi$ , is that it includes all five evolutionary events applied within current cophylogeny mapping algorithms including Jane, CoRe-PA and CoRe-ILP.

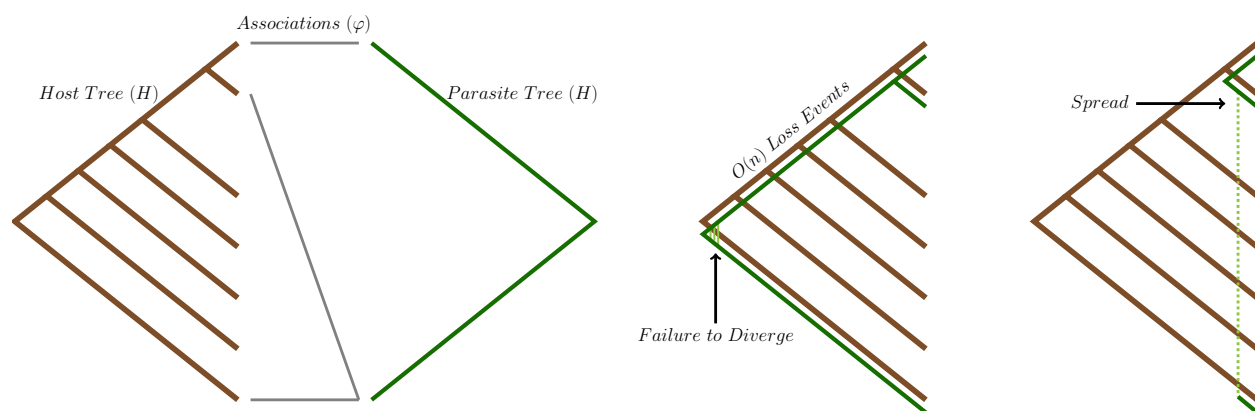


Figure 2: A tanglegram (left) and two Pareto optimal solutions using either failure-to-diverge (center) or spread (right).

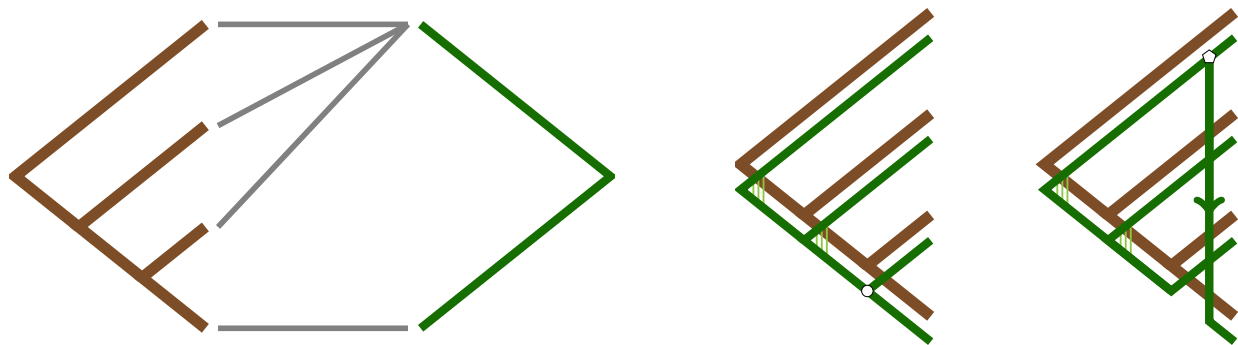


Figure 3: Tanglegram which will demonstrate why current implementations of widespread parasitism reconciliation using cophylogeny mapping fail (left), and two recovered maps which include the map recovered from Jane (right), and an optimal map (center). The algorithm presented herein is the first method proposed capable of presenting an algorithmic solution capable of recovering the optimal map for this tanglegram.

---

**Algorithm 1** RECONCILEWIDESPREADPARASITE( $H, P, \varphi, V, p$ )

---

```
1:  $\Phi$  is an array of lists which is worst case  $O(|P| \times |H|)$ 
2:  $p'$  is a homeomorphic sub-graph of  $h$  including the host leaves  $\in \varphi[p]$ 
3:  $L$  is a list of nodes in  $a$ 
4: Sort  $L$  in descending based on each nodes distance from the root of  $a$ 
5: for  $p'_i \in L$  do
6:   if  $p'_i$  is a leaf then
7:      $\Phi[p_i] \leftarrow$  leaf  $h_i \in H$  which  $p'_i$  is associated with
8:   else
9:      $l, r \leftarrow$  the left and right children of  $p'_i$ 
10:    for  $h_i \in \Phi[l]$  do
11:      for  $h_j \in \Phi[r]$  do
12:         $\Phi[p'_i][h_k] \leftarrow$  minimum cost event for  $p'_i$  at node  $h_k$ 
13:      end for
14:    end for
15:  end if
16: end for
17: return  $\Phi$ 
```

---

Figure 4: The ReconcileWideSpreadParasite subroutine called from the WiSPA algorithm (see Figure 5). This method outlines the process to infer the optimal set of widespread events from a set of association trees,  $A$ .

---

**Algorithm 2** WiSPA( $H, P, \varphi, V$ )

---

```
1:  $\Phi$  is an array of lists
2:  $\omega$  is an array of dynamic programming tables
3:  $L \leftarrow$  is a list of nodes in  $P$ 
4: Sort the nodes in  $L$  by their distance from the root of  $P$ 
5: for  $p_i \in L$  do
6:   if  $p_i$  is a leaf then
7:     if  $p_i$  is widespread then
8:        $\omega[p_i] \leftarrow \text{ReconcileWidespreadParasite}(H, P, \varphi, V, p_i)$ 
9:        $\Phi[p_i] \leftarrow \omega[p_i][p_i]$ 
10:    else
11:       $\Phi[p_i][h_i] \leftarrow$  leaf  $h_i \in H$  which  $p_i$  is associated with as defined in  $\varphi$ 
12:    end if
13:  else
14:     $l, r \leftarrow$  the left and right children of  $p_i$ 
15:    for  $h_l \in \Phi[l]$  do
16:      for  $h_r \in \Phi[r]$  do
17:        if  $h_l$  is a widespread mapping then
18:          for  $h'_l \in \text{AllFeasibleMappingSites}(H, P, \Phi, \Phi[h_l])$  do
19:             $h_p \leftarrow$  minimum cost mapping site for  $p_i$  for children  $h'_l$  and  $h_r$ 
20:             $\Phi[p_i][h_p] \leftarrow$  minimum cost event for  $p_i$  at node  $h_p$ 
21:          end for
22:        end if
23:        if  $h_r$  is a widespread mapping then
24:          for  $h'_r \in \text{AllFeasibleMappingSites}(H, P, \Phi, \Phi[h_r])$  do
25:             $h_p \leftarrow$  minimum cost mapping site for  $p_i$  for children  $h_l$  and  $h'_r$ 
26:             $\Phi[p_i][h_p] \leftarrow$  minimum cost event for  $p_i$  at node  $h_p$ 
27:          end for
28:        end if
29:        if  $h_l$  and  $h_r$  are not widespread mappings then
30:           $h_p \leftarrow$  minimum cost mapping site for  $p_i$  for children  $h_l$  and  $h_r$ 
31:           $\Phi[p_i][h_p] \leftarrow$  minimum cost event for  $p_i$  at node  $h_p$ 
32:        end if
33:      end for
34:    end for
35:  end if
36: end for
37: return  $\Phi(P)$ 
```

---

Figure 5: The WiSPA algorithm which outlines the process for reconciling the optimal set of widespread and divergence events for a pair of phylogenetic trees ( $H$  and  $P$ ), based on the known associations ( $\varphi$ ) between the two trees.

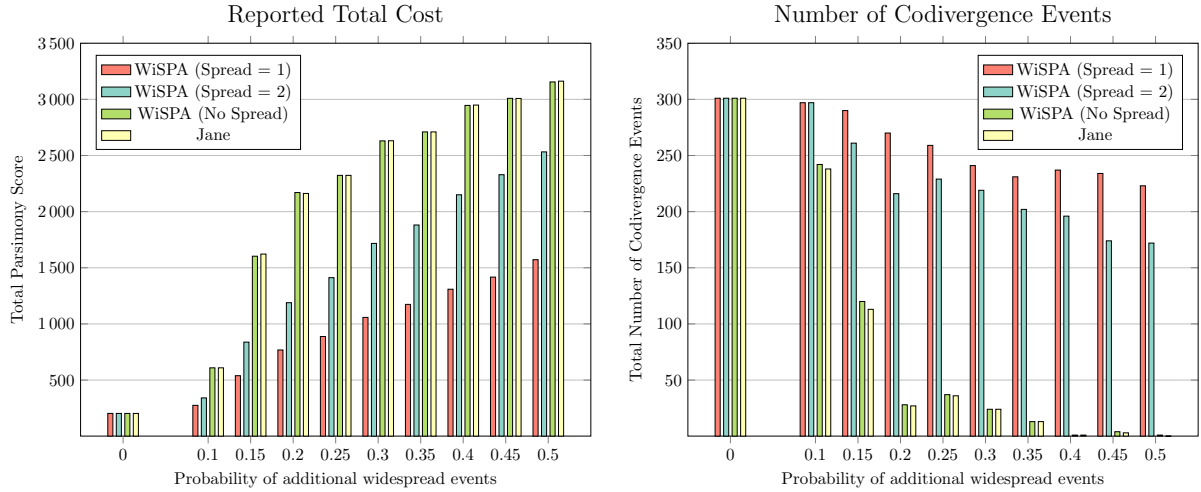


Figure 6: The results for the synthetic data sets. The first plot (left) considers the rate at which the total cost over 50 synthetic coevolutionary models increases as the rate of widespread parasitism is increased, where the second plot (right) considers the rate at which the total number of codivergence events over 50 synthetic coevolutionary models decreases as the rate of widespread parasitism is decreased.

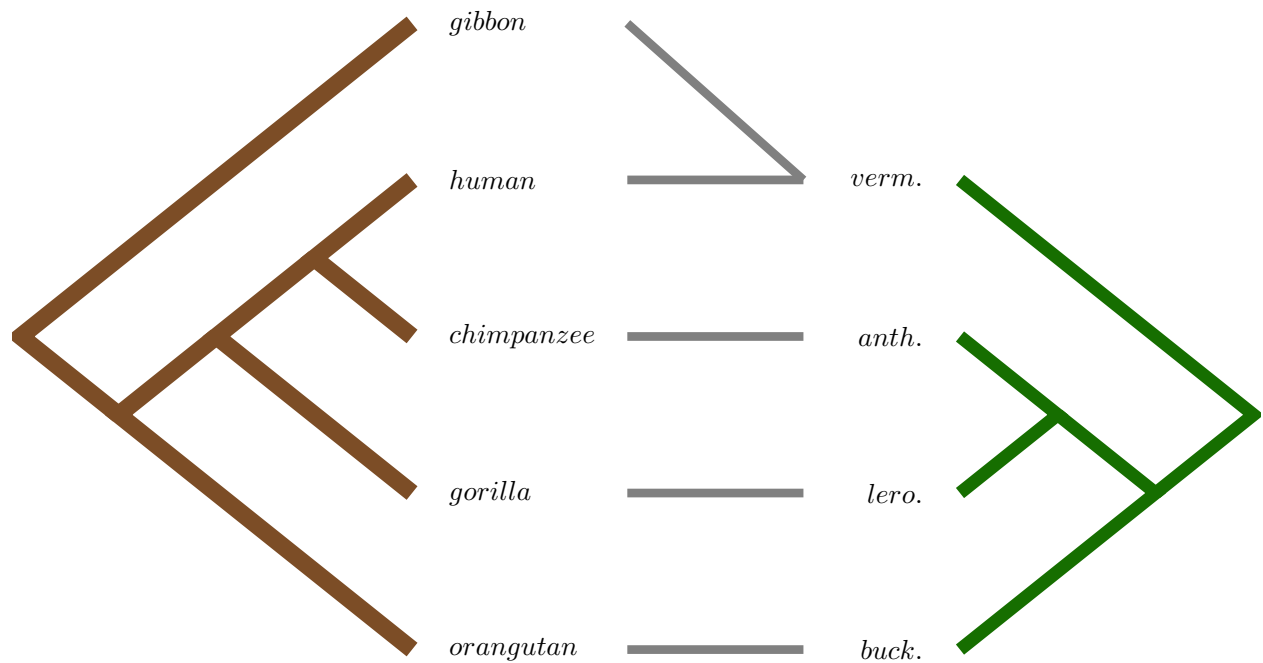


Figure 7: *Priamte-Enterobius* tanglegram adapted from Brooks and Glen (1982), and Ronquist (1997) where the widespread associations are marked in red.

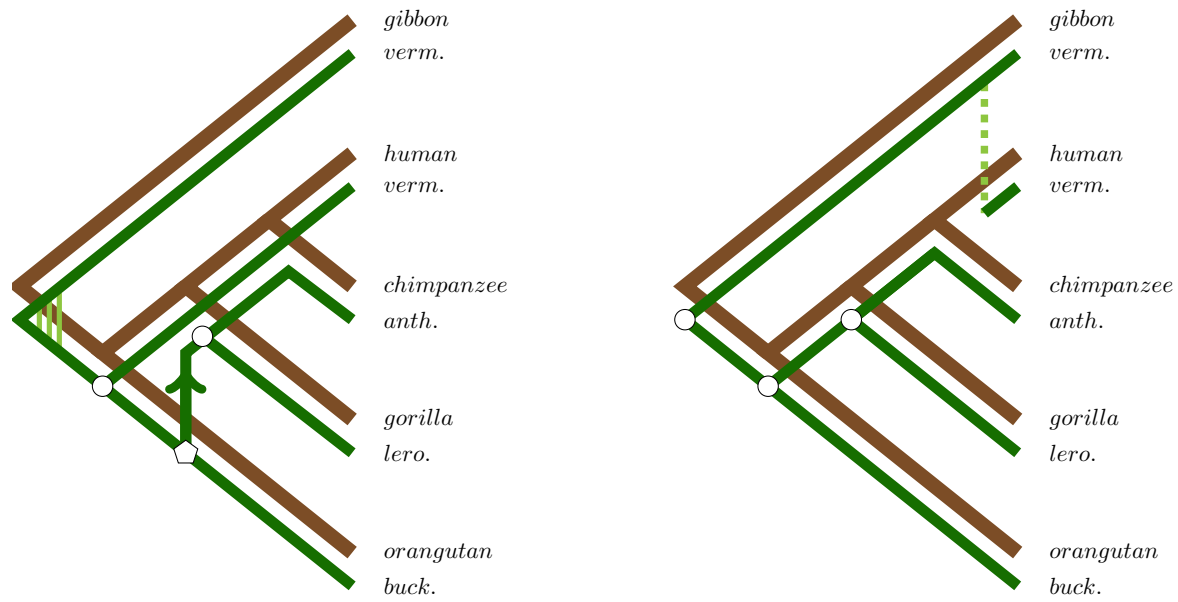


Figure 8: Two optimal maps recovered for the Primate / Pinworms data set. The first map (left) is the optimal reconstruction inferred using Jane, while the second map (right) is the optimal reconstruction inferred by WiSPA.



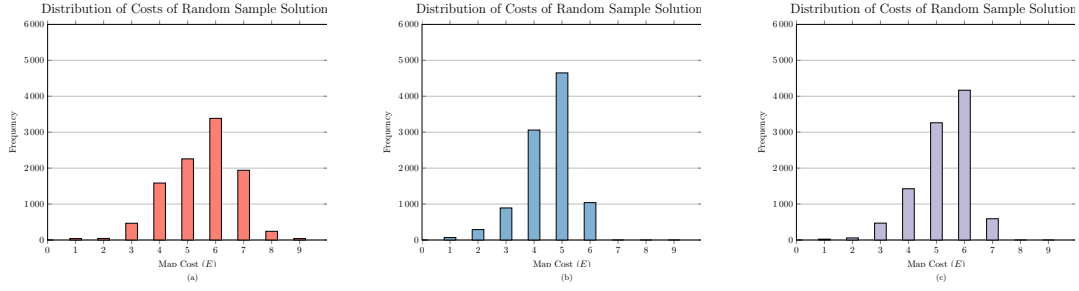


Figure 9: Results from the Bernoulli trials where 10000 replicates were run. Plot (left) records the distribution of the optimal reconstruction inferred using Jane while plot (center) and (right) record the distribution of the optimal reconstruction inferred by WiSPA for the cost scheme  $V = (0, 1, 2, 1, 1, 1)$  and  $V = (0, 1, 2, 1, 1, 2)$  respectively.